NPS62-85-027

NAVAL POSTGRADUATE SCHOOL Monterey, California





THE QUEUEING MODEL OF A PACKET SWITCH SUBJECT TO
ROUTING IN COMPUTER NETWORKS

JIN-FU CHANG

SEPTEMBER 1935

Approved for public release, distribution unlimited

Prepared for:

Chief of Naval Research Arlington, VA 22217

NAVAL POSTGRADUATE SCHOOL Monterey, California

Rear Admiral R. H. Shumaker Superintendent D. A. Schrady Provost

The Work reported herein was supported by the Foundation Research Program of the Naval Postgraduate School with funds provided by the Chief of Naval Research.

Reproduction of all or part of this report is authorized.

This report was prepared by:

JIN-FU CHANG

Associate Professor Electrical & Computer

Engineering

Reviewed by:

HARRIETT B. RIGAS

Chairman

Department of Electrical and

Computer Engineering

Released by:

JOHN DYÉR

Dean of Science and

Engineering

ではなける方法を表現なが、なるのなかない

REPORT DOCUMENTATION PAGE	READ INSTRUCTIONS BEFORE COMPLETING FORM	
1. REPORT NUMBER 2. GOVT ACCESSION NO.	1 DECIDIOUSIA CARALLAGA	
NPS62-85-027	31	
4. TITLE (and Subtitle)	S. TYPE OF REPORT & PERIOD COVERED	
THE QUEUEING MODEL OF A PACKET SWITCH SUBJECT	Technical	
TO ROUTING IN COMPUTER NETWORKS	6. PERFORMING ORG. REPORT NUMBER	
	S. CONTRACT OR GRANT NUMBER(s)	
7. AUTHOR(e)	61152N; RR000-01-NP	
Jin-Fu Chang	N0001485WR41005	
9. PERFORMING ORGANIZATION NAME AND ADDRESS	10. PROGRAM ELÉMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS	
Naval Postgraduate School Monterey, CA 93943-5100		
Honterey, CA 33343-3100		
11. CONTROLLING OFFICE NAME AND ADDRESS	12. REPORT DATE	
Chief of Naval Research	September 1985	
Arlington, VA 22217	13. NUMBER OF PAGES	
14. MONITORING AGENCY NAME & ADDRESS(II different from Controlling Office)	18. SECURITY CLASS, (of this report)	
	UNCLASSIFIED	
	150. DECLASSIFICATION/DOWNGRADING	
16. DISTRIBUTION STATEMENT (of this Report)	L	
Approved for public release; distribution unl	imited	
17. DISTRIBUTION STATEMENT (of the abstract entered in Black 20, if different fre	m Report)	
Approved for public release; distribution unlimited		
Approved for pastic release, distribution diffinited		
18. SUPPLEMENTARY NOTES		
19. KEY WORDS (Continue on reverse side if necessary and identify by block number)		
Computer Communication Networks, queue, routing, buffer sharing		
The effect of routing on the behavior of a finite queue accepting batch		
Poisson inputs is studied. Results obtained include state probabilities,		
blocking probabilities, delay and throughput. Results are also verified by		
computer simulations. The system considered is a good approximate model of		
a packet switch in a computer network.	5 3 7 7 7 7 1 1 1 1 1 1 1 1 1 1	

ABSTRACT

In a previous paper [1], the steady-state behavior of a finite queue which accepts batch poisson inputs and received service from servers operating in synchronous mode was studied. An analysis was successfully completed via the application of the Residue theorem in complex variables. This work extends the study in [1]-to include the effect of routing and buffer sharing. Upon the arrival of a batch each customer determines its route independently according to certain probability distribution. Buffer sharing with minimum allocation studied in [5] is also considered. Results obtained include state probability, blocking probability, delay, and throughput. Validity of analysis has been verified by computer simulations.

Accesio	n For	1
NTIS DTIC Unanno Justific	TAB ounced	
By Distribution /		
Availability Codes		
Dist	Avail and / o Special	or
A-1		



INTRODUCTION

In a previous paper [1] Chang and Chang study the steady state behavior of a queueing system under the following conditions.

- 1) Customers arrive at the queue in batches according to a Poisson process with mean rate λ batches/sec.
- 2) Each batch carries a random number of customers. The size of each batch is a positive integer valued random variable which may follow otherwise any arbitrary probability distribution.
 - 3) The system provides a capacity of accomodating N customers.
- 4) Customers receive services from m servers in the manner of first*come*first*served. Each customer requires a constant service time of T seconds. The servers operate synchronously in the sense that m customers can be removed form the system constantly every T seconds.
- 5) Upon the arrival of a batch if the remaining space is not enough to accept every customer in the batch then the entire batch will be completely rejected.

The system described above is an approximate model of a packet switch in a computer communication network. Although the problem is combinatorially very complex, it has been successfully solved via the application of the Residue theorem in complex variables. In a separate paper [2], an alternative approach called minislot approximation was introduced to solve the problem.

In a practical system such as computer communication network mentioned above, a packet switch may have several output channels and packets may have their own preferences in selecting an output channel. In order to make the model more practical and the results more valuable, part of the purpose of this study is to extend the work in [1] to include the effect of routing. In

other words, in addition to the above assumptions, should a batch be acceptable each customer of the batch decides from which server to receive service independently of the other customers according to a specific probability distribution. We shall use r_i to denote the probability that a customer will be routed to server i. Clearly, $0 \le r_i \le 1$ and $r_1 + \ldots + r_m = 1$. This type of routing is usually referred to as random routing in the area of computer communications.

The purpose of buffer sharing among customers routed to different outputchannels is to achieve efficient utilization of buffer. Several papers [3] ~ [5] have discussed this problem. It is pointed out in [5] that sharing with minimum allocation (SMA) performs better than complete sharing (CS) when traffic is high. In CS the entire buffer is accessible by all customers. In SMA each user has a reserved area which can be accessed only by customers routed to that server. In addition to the reserved region there is still a shared region. This work also tries to examine the effect of SMA on the behavior of a finite queue. For convenience we study the problem in which each server only has a reserved space capable of holding only one customer. Fig. 1 depicts the conceptual model of such a system. We believe the results can be modified to the situation where servers can have reserved areas of different sizes.

The approach using the Residue theorem proposed in [1] will again be used to obtain results such as state probability, blocking probability, average delay, and system throughput. However this approach has to be modified. The modification is necessitated by routing consideration. For example the derivation of state transition probability via the Residue theorem not only depends on the remaining buffer space but also on the set of busy and idle servers. This will be demonstrated in Section 2.

Section 3 of this report discusses the extension of this work to unequalrate servers. Conclusions are made in Section 4.

Throughout this work ergodicity of the process is assumed so that essemble—averages can be replaced by time—averages.

2. MAIN RESULTS

Let π_b denote the probability that the system is at state b = (b₁, ..., b_m) at the beginning of a time slot. In $b = (b_1, \ldots, b_m)$ such that $b = b_1 + b_1$... + b_m , $b_i(\geq 0)$ represents the number of customers which are to be routed to server i while b denotes the total number of customers in the system. Let x = (x_1, \ldots, x_m) such that each x_i is nonnegative integer and $x_1 + \ldots + x_m > 0$ be an arriving batch of size $x_1 + ... + x_m$ in which x_i of them are to be routed to server i. Suppose N = 10, m = 2, and h = (1,2). Also suppose three arrivals (3,1), (2,6), (1,1) (according to their order of appearance) have occurred within a slot. The first arrival (3,1) is acceptable and will lead the system state from (1,2) to (4,3). The second arrival (2,6) must be rejected since 2 + 6 + 4 + 3 = 15 > 10. The last arrival (1,1) is acceptable since 1 + 1 + 4 + 3 = 9 > 10. In other words if b = (1,2), the arrivals (3,1), (2,6), (1,1) all together result in four and two customers effectively accepted by the system and to be routed to server 1 and 2, respectively. We use a = (4,2) = (3,1) + (1,1) to illustrate such an event and use $p_{a,b}$ to denote the corresponding conditional probability. In general, \underline{a} = $(a_1, \ldots,$ a_m) in which each a_i is a nonnegative integer such that $a = a_1 + ... + a_m > 0$.

Let $n = (n_1, \ldots, n_m)$ denote the system state at the beginning of the next slot. Then the steady-state behavior of the system is characterized by

(1)
$$\pi = \sum_{b \in A} \pi_b p_{a,b}$$

 $\underline{n} (\underline{a}, \underline{b}) \in AB(\underline{n})$

where

(2)
$$AB(\frac{n}{2}) = \{(a, b) \mid 0 \le (b_i - 1)^+ < n_i,$$

$$\sum_{i=1}^{m} (b_{i} - 1)^{+} \leq N - m,$$

$$a_{i} = n_{i} - (b_{i} - 1)^{+},$$

$$a + b \leq N,$$
and
$$\sum_{i=1}^{m} (a_{i} + b_{i} - 1)^{+} \leq N - m$$

In (2),

(3)
$$(x)^{+} = \begin{cases} x & \text{if } x > 0 \\ 0 & \text{if } x \leq 0 \end{cases}$$

Physically, AB(\underline{n}) represents the set of (a, b) such that the system can transit from \underline{b} to \underline{n} via the acceptance of \underline{a} . In (2), $(b_i - 1)^+$ represents the number of customers (among the initial b_i) remain in the system at the end of the slot. $\sum_{i=1}^m (b_i - 1)^+$ must not be greater than N-m since at most m

customers can be removed from the system per slot. $\sum_{i=1}^{m} (b_i + 1)^+ = N + m$

occurs when b = N, i.e. system is full, at the beginning of a slot and all servers are busy during the slot. $(a_i + b_i + 1)^+$ represents the number of customers which stay in the shared region and to be routed to server i. For example if $b_i = 0$ and $a_i = 1$, then this customer will be placed in the reserved area and there is no customer waiting for server i in the shared area.

Since the shared area is of size N=m, we have $\sum_{i=1}^{m} (a_i + b_i = 1)^+ \le N - m$ in (2).

 $\boldsymbol{\pi}_{b}$ in (1) must be solved together with

$$\begin{array}{ccc} \begin{array}{ccc} (1) & & & \sum & \pi_{\underline{b}} & = & 1 \\ & b & \epsilon & BB & & \underline{b} \end{array}$$

where

(5) BB = {
$$b \mid 0 \le b_i \le N = (m + 1), 1 \le i \le m, 0 \le b \le N }$$

Physically, BB denotes the set of feasible states. In (5), N \neq (m - 1) represents that the system can have at most N = (m - 1) customers to be routed to server i, 1 \leq i \leq m.

In order to solve (1), we need to find p_a , p_i first. Let x_i be a sequence of batch arrivals arranged in the order of their appearances in which x_i = $(x_{i1}, x_{i2}, \ldots, x_{im})$. We have explained previously that x_{ij} represents the number of customers in the ith batch of $\{x_i\}$ which select server j to receive their services. Next we use B and I to denote the set of busy and idle servers, respectively within a slot. A customer which arrives within a slot must wait till the beginning of the next slot to receive service even if the server it selects is idle when it arrives. This is because the servers are running synchronously at the same speed. The duration between the arrival of a customer and the beginning of the next slot will be called residual period of this customer. Let R denote the size of the remaining free buffer

including both the shared and reserved region at the beginning of time slot. Let $\mid B \mid$ and $\mid I \mid$ respectively denote the number of busy and idle servers. Clearly, $R > \mid I \mid$ and $R - \mid I \mid$ represents the size of free buffer in the shared area.

The approach using the residue theorem proposed in [1] can be used to obtain p_a , b. However it has to be modified.

Example 1. Suppose R = 5, m = 3, B = $\{2\}$, and I = $\{1, 3\}$. This means the shared region can take no more than 3 customers. Thus in $\{x_1\}$ = $\{(4,3,0), (0,3,1), (1,4,2), (5,1,2), (0,0,1), (0,3,0)\}$ only (0,3,1) is acceptable and the rest are rejected. The arrival (0,0,1) is rejected since after the acceptance of (0,3,1) not only the shared region becomes full but also the buffer reserved for server 3 is occupied. Thus the only customer in (0,0,1) which can only be placed in the shared region gets rejected. $\{(0,3,1)\}$ is called an acceptable pattern when R = 5, B = $\{2\}$, and I = $\{1, 3\}$. $\{(0,3,1)\}$ further divides $\{x_1\}$ into two subsequences $C_0 = \{(4, 3, 0)\}$ and $C_1 = \{(1, 4, 2), (5,1,3), (0,0,1), (0,3,0)\}$. C_0 and C_1 will be called blocked subsequences of $\{x_1\}$.

In general, we use $\{a_1, \ldots, a_n\}$ to denote a general acceptable pattern such that $\frac{a}{k} = (a_{k1}, a_{k2}, \ldots, a_{km})$ and $a_k = a_{k1} + \ldots + a_{km}$. Let v_i and u_i respectively denote the number of unoccupied reserved buffers and the number of unoccupied shared buffers before the acceptance of $\underline{a_{i+1}}$, $i = 0, 1, \ldots, n+1$. For i = n, u_n and v_n are defined similarly except after the acceptance of $\underline{a_n}$. Use R_i to denote the size of the remaining free shared and reserved buffers between the acceptances of $\underline{a_i}$ and $\underline{a_{i+1}}$. Then any arrival pattern $\{\underline{x_i}\}$ which contains $\underline{a_1}, \ldots, \underline{a_n}$ as a subset may have the following blocked subsequences

 \underline{c}_0 , \underline{c}_1 , ..., \underline{c}_n such that $c_{ij} \in \underline{c}_i$, $c_{ij} = (c_{ij1} \dots c_{ijm})$ with $c_{ij} = c_{ij1} + \dots + c_{ijm}$ satisfies

- (6.a) $c_{oj} > R_o = R$ or the number of unoccupied reserved buffers selected by customers specified in c_{oj} is less than $c_{oj} v_o$ for $v_o + 1 \le c_{oj} \le v_o + v_o = v_o$
- (6.b) $c_{ij} > R_j = R = \sum_{k=1}^{i} a_k$ or the number of unoccupied reserved buffers

selected by customers specified in $\underline{c_{ij}}$ is less than $c_{ij} = v_i$ for $v_i + 1 \le c_{ij} \le c_{ij} \le u_i + v_i = R_i$, $i = 1, \ldots, n$.

Define for i = 1, ..., n and after the acceptance of $\underline{a_i}$

- (7) $B_i = \{j \mid \text{server j for which the reserved buffer is occupied by some customer} \}$
- (8) $I_i = \{j \mid \text{server } j \text{ idle and its reserved buffer is still empty}\}$ Clearly, $B_i \cup I_i = \{1, 2, ..., m\}$, $u_i = |I_i|$, and $v_i = |B_i|$. Next, let $B_0 = B$ and $I_0 = I$ where B and I have been defined previously to be the number of busy and idle servers, respectively, at the beginning of a slot. Also define q_j (x, R_i , B_i , I_i) to be the probability that an arriving batch of size x is rejected because the number of servers in I_i selected by customers in the batch is j which is less than $x = v_i$. For example

(9)
$$q_{i}(x, R_{i}, B_{i}, I_{i}) = (\sum_{\ell \in B_{i}} r_{\ell}), x \ge v_{i} + 1$$

while

(10)
$$q_1(x, R_i, B_i, I_i) = \sum_{k \in I_i} \left[\left(\sum_{k \in B_i} r_{\ell_{\{k\}}} \right)^x - \left(\sum_{k \in B_i} r_{\ell} \right)^x \right]$$

$$= \sum_{k \in I_{i}} [(\begin{array}{ccc} r \\ \ell \in B_{i} \end{array})^{x} \sim q_{o}(x, R_{i}, B_{i}, \{k\})], x \geq v_{i} + 2$$

In general,

(11)
$$q_{j}(x, R_{i}, B_{i}, I_{i}) = \sum_{(k_{1}, ..., k_{j})} [\sum_{\epsilon KK_{j}} (\sum_{\epsilon BU\{k_{1}, ..., k_{j}\}})^{x}$$

$$= \sum_{k=0}^{j=1} q_k (x, R_i, B_i, \{k_1, \dots k_j\})], x \ge v_i + j + 1$$

where

(12)
$$KK_j = \{k_1, ..., k_j\}$$
 each $k_l \in I_i$ and $k_1, ..., k_j$ distinct

Eq. (11) above can be easily encoded into computer program., Our expreiences show the enumeration of $q_j(x, R_i, B_i, I_i)$ does not require too much computer time.

Let β_i denote the probability that an arriving batch is rejected between the acceptance of \underline{a}_i and \underline{a}_{i+1} for $i=1,2,\ldots,n-1$. β_0 and β_n resepctively denote the above probability before the acceptance of \underline{a}_1 and after the acceptance of \underline{a}_n . Then from (6.a) and (6.b)

(13)
$$\beta_{0} = \overline{g}_{R_{0}} + \sum_{x = v_{0} + 1}^{R_{0}} g_{x} \left[\sum_{j = 0}^{x - v_{0} - 1} g_{j}(x, R_{0}, B_{0}, I_{0}) \right]$$

where g_{X} denotes the probability that a batch contains a total of x customers and

(14)
$$\ddot{g}_{x} = g_{x+1} + g_{x+2} + ...$$

represents the probability that a batch c ntains more than x customers. Next

(15)
$$\beta_1 = \tilde{g}_{R_1} + \sum_{x=v_1+1}^{R_1} g_x \left[\sum_{j=0}^{x-v_1-1} q_j (x, R_1, B_1, I_1) \right]$$

In general

(16)
$$\beta_{i} = \overline{g}_{R_{1}} + \sum_{x=v_{i}+1}^{R_{i}} g_{x} \left[\sum_{j=0}^{x} -1 q_{j} (x, R_{i}, B_{i}, I_{i}) \right]$$

Let $N_{\rm a}$ be the random variable representing the number of batches arrive in a slot and let AP stand for acceptance pattern then

(17)
$$P[AP = \{\underline{a}_1, \underline{a}_2, ..., \underline{a}_n\} | N_a = k]$$

$$= \begin{cases} \prod_{i=1}^{n} [g_{a_{i}} (\begin{array}{c} a_{1} \\ \end{array}) \dots (\begin{array}{c} a_{1} \\ \end{array} \sum_{i=1}^{m-1} a_{ij}) \prod_{i=1}^{m} (r_{j})^{a_{ij}} \sum_{j=1}^{B_{0}^{j0}} B_{1}^{j1} B_{n}^{jn}, \text{ if k n} \\ j_{0}^{+} \dots + j_{n} = k^{a_{n}} \\ 0, \text{ if } k < 0 \end{cases}$$

where $a_1+\ldots+a_n\leq R$. Via the approach using the Residue theorem [1] we obtain

(18)
$$P[AP = \{\underline{a}_1, \underline{a}_2, \dots, \underline{a}_n\} \mid N_a = k]$$

$$= CAP \frac{1}{2\pi i} \oint \left| z \right| = c \frac{z^k}{\prod_{j=0}^{n} (z - \beta_j)} dz$$

where $c > \beta_0, \beta_1, \ldots, \beta_n$ and

(throughout this report the i assoicated with $2\pi i$ denote $\sqrt{-1}$, otherwise it can be used as a subscript or superscript.) Finally,

(20)
$$P[AP = \{\underline{a}_1, \underline{a}_2, \dots, \underline{a}_n\}]$$

$$= \sum_{k=0}^{\infty} P[AP = \{a_1, ..., a_n\} \mid_{N_a = k]} f_k$$

$$= \frac{\text{CAP}}{2\pi i} \oint |z| = c \frac{F(z)}{\prod_{j=0}^{n} (z - \beta_j)} dz$$

where $f_k = P[N_a = k]$ and $F(z) = f_0 + f_1 z + f_2 z^2 + ...$ is the probability generating function of f_k . The derivation of (20) is similar to that in [1] where routing is not considered.

Define

(21)
$$\Psi_a = \Psi (a_1 a_2 ... a_m)$$

= $\{ AP = \{ \underline{e_1}, ..., \underline{e_n} \} | e_i = (e_{i1}, e_{i1}, ... e_{im}), \}$

$$\sum_{k=1}^{n} e_{kj} = a_{j}, 1 \leq j \leq m, 1 \leq n \leq R$$

We use $|\Psi_a|$ to denote the number of APs contained in Ψ_a . Similar to [1], Ψ_a can be obtained recursively as follows. If AP = $\{\underline{e}_1, \ldots, \underline{e}_n\}$ then define (22) AP * \underline{e}_{n+1} = $\{\underline{e}_1, \ldots, \underline{e}_n, \underline{e}_{n+1}\}$

Let $\Psi = \{AP_1, \dots, AP_\ell\}$ where ℓ is an arbitrary positive integer and e an arbitrary acceptable batch, then define

(23)
$$\Psi * g = \{ AP_1 * g, ..., AP_k * g \}$$

The definitions of (22) and (23) are taken directly from [1].

$$\begin{split} \Psi_{(1,0)} &= \{\{(1,0)\}\} \\ \Psi_{(0,1)} &= \{\{(0,1)\}\} \\ \Psi_{(2,0)} &= \{\{(2,0)\}, \{(1,0), (1,0)\}\} \\ &= \{\Psi_{(0,0)} * (2,0)\} \quad \{\Psi_{(0,1)} * (1,0)\} \} \\ \Psi_{(1,1)} &= \{\{(1,1)\}, \{(1,0), (0,1)\}, \{(0,1), (1,0)\}\} \\ &= \{\Psi_{(0,0)} * (1,1)\} \quad \{\Psi_{(1,0)} * (0,1)\} \quad \{\Psi_{(0,1)} * (1,0)\} \} \\ \Psi_{(2,1)} &= \{\{(2,1)\}\} \quad \{\{(2,0), (0,1)\}, \{(1,0), (1,0), (0,1)\} \} \\ &= \{\{(1,1), (1,0)\}, \{(1,0), (0,1), (1,0)\}, \{(0,1), (1,0), (1,0)\} \} \\ &= \{\Psi_{(0,0)} * (2,1)\} \quad \{\Psi_{(2,0)} * (0,1)\} \quad \{\Psi_{(1,1)} * (1,0)\} \\ &= \{\Psi_{(0,1)} * (2,0)\} \quad \{\Psi_{(1,0)} * (1,1)\} \end{split}$$

In general, we have

$$\frac{\Psi_{\underline{a}} = \Psi(a_1, \ldots, a_m)}{\Psi(a_1, \ldots, a_m)}$$

$$\{ \psi_{(a_1, \ldots, k_m) \in \Phi_{(a_1, \ldots, a_m)}}^{(k_1, \ldots, k_m) \in \Phi_{(a_1, \ldots, a_m)}^{(k_1, \ldots, k_m)}}^{(k_1, \ldots, k_m) \in \Phi_{(a_1, \ldots, a_m)}^{(k_1, \ldots, k_m) \in \Phi_{(a_1, \ldots, a_m)}^{(k_1, \ldots, k_m) \in \Phi_{(a_1, \ldots, a_m)}^{(k_1, \ldots, k_m)}}^{(k_1, \ldots, k_m) \in \Phi_{(a_1, \ldots, a_m)}^{(k_1, \ldots, k_m)}}^{(k_1, \ldots, k_$$

where

(25)
$$\Phi(a_1, \ldots, a_m) = \{(k_1, \ldots, k_m) \mid \le k_1 \le a_1, \ldots, 0 \le k_m \le a_m, \text{ but } (k_1, \ldots, k_m) \neq (0, \ldots, 0)\}$$

Now let us find $P[\Psi_a]$.

Example 3. Let m = 2. From our previous discussion of (13), (15) \leftarrow (16) we have

(26)
$$P[\Psi_{(0,0)}] = \frac{1}{2\pi i} \oint \frac{F(z)}{(z - \beta_{00})} dz$$

$$= \frac{1}{2\pi i} \oint \frac{A_{00}(z) F(z)}{B_{00}(z)} dz$$

where A_{00} (z) = 1, B_{00} (z) = z - β_{00} , and β_{00} can be obtained from (13) assuming R_0 , B_0 and I_0 are known.

If $(a_1, a_2) = (1,0)$ we have

(27)
$$P[\Psi_{(1,0)}] = \frac{1}{2\pi i} \oint \frac{g_1 r_1 F(z)}{(z - \beta_{00}) (z - \beta_{10})} dz$$
$$= \frac{1}{2\pi i} \oint \frac{A_{10} (z) F(z)}{B_{10} (z)} dz$$

where $A_{10}(z) = g_1r_1$, $B_{10}(z) = (z - \beta_{00})(z - \beta_{10})$. In (27), β_{00} and β_{10} can be obtained from (13) and (15) under the assumption that R_0 , B_0 , I_0 are known and $\{(1,0)\}$ is only acceptable pattern.

Similarly if $(a_1, a_2) = (0,1)$ we have

(28)
$$P[\Psi_{(0,1)}] = \frac{1}{2\pi i} \oint \frac{A_{01}(z) F(z)}{B_{01}(z)} dz$$

where $A_{01}(z) = g_1r_1$, $B_{01}(z) = (z - \beta_{00})(z - \beta_{01})$. In (28), β_{00} and β_{01} can be obtained from (13) and (15) assuming R_0 , B_0 , I_0 are known and $\{(0,1)\}$ is the only acceptable pattern. If $(a_1, a_2) = (2,0)$, since $\Psi_{(2,0)} = \{\{(2,0)\}\}$ U $\{\Psi_{(1,0)}^* = \{(1,0)\}\}$ we have

(29)
$$P[\Psi_{(2,0)}] = P[AP = \{(2,0)\}] + P[AP = \{(1,0), (1,0)\}]$$

$$= \frac{1}{2\pi i} \oint \frac{A_{20}(z) F(z)}{B_{20}(z)} dz$$

where

$$A_{20}(z) = g_2 r_1^2 A_{00}(z) \frac{B_{20}(z)}{B_{00}(z)(z - \beta_{00})} + \frac{g_1 r_1 A_{10}(z) B_{20}(z)}{B_{10}(z)(z - \beta_{20})}$$

$$= \sum_{(k_1, k_2) \in \Phi_{(2,0)} k_1 + k_2} r_1^{k_1} r_2^{k_2} {k_1 + k_2 \choose 1} \frac{A_{2-k_1,0} B_{20}(z)}{B_{2-k_1,0}(z)(z + \beta_{20})}$$

$$B_{20}(z) = (z - \beta_{00})(z - \beta_{10})(z - \beta_{20})$$

$$= 0 \quad (z - \beta_{10})(z - \beta_{20})$$

$$= 0 \quad (z - \beta_{10})(z - \beta_{10})(z - \beta_{20})$$

Suppose R₀, B₀, and I₀ are given, β_{20} can be obtained either from β_{1} of (15) by assuming AP = $\{(2,0)\}$ or from β_{2} of (16) assuming AP = $\{(1,0), (1,0)\}$.

In general, for m = 2, we have

(30)
$$P[\Psi(a_1, a_2)] = \frac{1}{2\pi i} \oint \frac{A_{a_1, a_2}(z) F(z)}{B_{a_1, a_2}(z)} dz$$

where

(32)
$$B(a_1, a_2) = \prod_{\beta_{k_1, k_2}} (z^{+\beta_{k_1, k_2}})$$

In (32), BTA $(a_1, a_2) = \{\beta_{k_1, k_2} \mid 0 \le k_1 \le a_1, 0 \le k_2 \le a_1, \text{ and all } \beta_{k_1, k_2} \}$ must have distinct values. In other words, there could be several combinations of k_1 and k_2 so that their values of β are identical. In this

case, we select only one of them as representative. This can be explained as follows.

In the process of combining all the related rationals to reach

$$A_{a_1}$$
, A_{a_2} A_{a_1} , A_{a_2} $A_$

$$\beta_{k_1, k_2} = \beta_{k_1, k_2}$$
. In this case we only let the factor $(z - \beta_{k_1, k_2})$ appear

once in B_{a_1, a_2} (z). Thus, BTA (a_1, a_2) used in (32) and defined in the line

following (32) is for this purpose. Notice that $\phi(a_1, a_2)$ has been defined in

(25). In practice, the value of
$$\beta_{k_1, k_2}$$
 which appears in β_{a_1, a_2} (z) can be

obtained from β_1 of (15) by assuming R_0 , B_0 , I_0 are given and $AP = \{(k_1, k_2)\}$. Finally.

(33)
$$p_{\underline{a}}, \underline{b} = P(a_1, a_2), \underline{b}$$

=
$$P[\Psi(a_1, a_2) | R_0 = N = b, B_0 = B, I_0 = I]$$

Throughout this example, all the integrals are assumed to be carrried out along a circle whose radius is bigger than any of the poles.

(34)
$$P[\Psi_{\mathbf{a}}] = P[\Psi_{(\mathbf{a}_{1}, \ldots, \mathbf{a}_{m})}]$$

$$= \frac{1}{2\pi i} \int \sum_{\underline{k} \in \Phi_{\mathbf{a}}} g_{k_{1}} + \ldots + k_{m} {k_{1} + \ldots + k_{m} \choose k_{1}} {k_{2} + \ldots + k_{m} \choose k_{m}} {k_{2} + \ldots + k_{m} \choose k_{m}}$$

$$\frac{r_1^{k_1} r_2^{k_2} \dots r_m^{k_m} A_{a_1 - k_1, \dots, a_m + k_m}(z) F(z)}{B_{a_1 - k_1, \dots, a_m + k_m}(z) (z + \beta_{a_1 a_2 \dots a_m})} dz$$

$$= \frac{1}{2\pi i} \int \frac{A_{a_1}, \dots, a_m(z)F(z)}{B_{a_1}, \dots, a_m(z)} dz$$

where $\underline{K} = (k_1, \ldots, k_m)$ and

(35)
$$A_{a_1}, \dots a_m$$
 (z) = $\sum_{\underline{k} \in \Psi_a} g_{k_1 + \dots + k_m} \binom{k_1 + \dots + k_m}{k_1} \binom{k_2 + \dots + k_m}{k_2} \binom{k_m}{k_m}$

$$\frac{r_{1}^{k_{1}}...r_{m}^{k_{m}}A_{a_{1}-k_{1}},...a_{m}=k_{m}(z)\cdot B_{a_{1}},...,a_{m}(z)]}{[B_{a_{1}}-k_{1},...,a_{m}=k_{m}(z)(z-\beta_{a_{1}},...a_{m})]}$$

(36)
$$B_{a_{1}}, \dots a_{m} \qquad (z) = \prod_{\substack{\beta \\ k_{1}, \dots k_{m} \\ }} (z - \beta_{k_{1}}, \dots, k_{m})$$

In (36),

(37) BTA
$$(a_1, \ldots, a_m) = \{\beta_{k_1}, \ldots, k_m \mid 0 \le k_i \le a_i \text{ and all } \beta_{k_1} \ldots k_m \}$$
must have distinct values

Also, β_{k_1} , ..., k_m can be obtained from (15) by assuming R_0 , B_0 , I_0 are known

and AP = $\{(k_1, ..., k_m)\}$, i.e.,

(38)
$$\beta_{k_1}, \ldots, k_m = g_{R_1} + \sum_{x = v_1 + 1}^{R_1} g_x \begin{bmatrix} x & y_1 - 1 \\ y_1 & y_1 \end{bmatrix} q_j (x, R_1, B_1, I_1) \end{bmatrix}$$

where

(39a)
$$R_1 = R_0 - \sum_{j=1}^{m} k_j$$

(39b)
$$B_1 = \{B_0 \cup \{1 \cdot u(k_1), 2 \cdot u(k_2), ..., m \cdot u(k_m)\} \cup \{0\}\} - \{0\}$$

(39c)
$$I_1 = I_0 - \{1 \cdot u(k_1), 2 \cdot u(k_2), ..., m \cdot u(k_m)\}$$

(39d)
$$v_1 = R_1 = I_1$$

In (39), u(x) denotes the unit-step function, i.e.,

(40)
$$u(x) = \begin{cases} 1 & \text{if } x > 0 \\ 0 & \text{if } x < 0 \end{cases}$$

Notice that $\{0\}$ and $-\{0\}$ in B_1 is to prevent from including 0 and B_1 since servers are numbered from 1 to m.

Finally, p_a , b can be evaluated as follows

(41)
$$p_{a,b} = P[\Psi_a \mid B_0 = B, I_0 = I, R_0 = N - b]$$

This is because at the beginning of a slot, the initial state \underline{b} is completely specified by the set of busy servers B, the set of idle servers I, and the size of the remaining free buffer space N-b. To evaluate the probability that \underline{a} is accepted is equivalent to evaluate $\Psi_{\underline{a}}$ based on B, I and N=b. Once $p_{\underline{a},\underline{b}}$ is obtained, (1) can be solved. The finiteness of the queue guarantees the existence of the probabilities $\pi_{\underline{n}}$ regardless of the value of λ . The minislot approximation porposed in [2] can also be used to obtain $p_{\underline{a},\underline{b}}$.

Now let $\pi^t_{\underline{n}}$ denote the probability that the system is at state \underline{n} at t seconds after the beginning of a slot. clearly π^t_n satisfies

(42)
$$\pi_n^t = \sum_{(\underline{a}, \underline{b})} \pi_{\underline{b}} p_{\underline{a}, \underline{b}}^t$$

subject to

$$(43) \qquad \sum_{\underline{n} \in BB} \qquad \pi_{\underline{n}}^{t} = 1$$

where ABT(n) = $\{(\underline{a}, \underline{b}) \mid b_j \leq n_j \text{ and } a_j = n_j + b_j\}$ and BB has been defined in

(5). Also in (42) $p_{\underline{a}, \ \underline{b}}^t$ is the probability \underline{a} has been accepted into the system if at the beginning of the slot the system state was \underline{b} . Let 0 denote the beginning of a slot and define

Then

(45)
$$\pi_{\underline{n}}^* = \sum_{(\underline{a}, \underline{b}) \in ABT(\underline{n})}^* \pi_{\underline{p}} p_{\underline{a}, \underline{b}}^*$$

subject to

$$\sum_{\underline{n} \in BB} \pi_{\underline{n}}^* = 1$$

In (45)

(47)
$$p_{\underline{a},\underline{b}}^* = \frac{1}{T} \int_0^T p_{\underline{a},\underline{b}}^t dt$$

$$= \frac{1}{2\pi i} \frac{1}{T} \int_{0}^{T} \frac{A_{\underline{a}}(z)F^{t}(z)}{B_{\underline{a}}(z)} dz dt$$

$$= \frac{1}{2\pi i} \frac{1}{T} \int \frac{A_a(z)}{B_a(z)} \int_0^T F^t(z) dt dz$$

$$= \frac{1}{2\pi i} \frac{1}{T} \int \frac{A_a(z)}{B_a(z)} \int_0^T e^{-\lambda t(1-z)} dt dz$$

$$= \frac{1}{\lambda T} \frac{1}{2\pi i} \oint \frac{A_{a_1}, \dots, a_m}{B_{a_1}, \dots, a_m} (z) \left[e^{\lambda T(z^{\perp}1)} - 1 \right] dz$$

$$= \frac{1}{\lambda T} \frac{1}{2\pi i} \oint \frac{A_{a_1}, \dots, a_m}{B_{a_1}, \dots, a_m} (z) (z^{\perp}1) dz$$

The derivation of (47) is similar to the derivation of $p_{a,b}$.

We use B_b^t to denote the probability that a batch arrives at t (0,T) is rejected. Also we use B_b to denote the average blocking probability of a batch. Clearly,

(48)
$$B_b = \frac{1}{T} \int_0^T B_b^t dt$$

$$= \sum_{n \in BB} \pi_n^* \left\{ g_{N-n}^* + \sum_{x=v(n)+1}^{R(n)} g_x^* [\sum_{j=0}^{x=v(n)-1} q_j^*(x, R(n), B(n), I(n))] \right\}$$

where

(49a)
$$B(\underline{n}) = \{\{1 \cdot u(n_1), 2 \cdot u(n_2), ..., m \cdot u(n_m)\} \cup \{0\}\} = \{0\}$$

(49b)
$$1(\underline{n}) = \{1, ..., m\} + B(\underline{n})$$

(49c)
$$R(\underline{n}) = N = b$$

(49d)
$$v(\underline{n}) = R(\underline{n}) - |I(n)|$$

Define a test customer to be a randomly selected customer. Clearly, the probability that this test customer is drawn from a batch of size x is $xg_X/$

 $(g_1 + 2g_2 + \dots)$. Let B_c^t be the blocking probability of a test customer t seconds after the beginning of a slot and

$$B_{c} = \frac{1}{T} \int_{0}^{T} B_{c}^{t} dt,$$

then

(50)
$$B_{c} = \frac{1}{a} \sum_{n=0}^{N} \pi_{n}^{*} \left\{ \sum_{x=N-n+1}^{\infty} xg_{x} + \sum_{x=v(n)+1}^{R(n)} xg_{x} \right\}$$

$$\left[\sum_{j=0}^{x-v(n)-1} q_{j}(x, R(n), B(n), I(n)) \right]$$

where

(51)
$$\alpha = \sum_{X=1}^{\infty} xg_X$$

represents the average batch size.

Let $\boldsymbol{L}_{\boldsymbol{Q}}$ denote the average number of customers waiting in the system, then

(52)
$$L_{q} = \sum_{\underline{b} \in BB} \left[\left(\sum_{\underline{a} \in AA(b)}^{\underline{m}} a_{\underline{j}} + (b_{\underline{j}} - 1)^{+} \right) p_{\underline{a},\underline{b}}^{\underline{n}} \right] \Pi_{\underline{b}}$$

where $AA(\underline{b})$ denotes the set of feasible \underline{a} if the system was initially at \underline{b} and can be expressed as follows

(53)
$$AA(\underline{b}) = \{\underline{a} \mid 0 \le a \le N - b, 0 \le \sum_{j=1}^{m} (a_j + b_j - 1)^+ \le N - m\}$$

The reason that $\sum_{j=1}^{m} (a_j + b_j - 1)^+ \le N - m$ has been given in the explanation

of (2). Next let $L_{\mbox{\scriptsize q}\,,\,\mbox{\scriptsize k}}$ denote the average queue length associated with server k, then

(54)
$$L_{q,k} = \sum_{b \in BB} \left[\sum_{a \in AA(k)} (a_k + (b_{k-1})^+) P_{a,b}^* \right] \prod_{\underline{b}}$$

We use $\lambda_{e,k}$ to denote the effective input rate, expressed in number of customers per sec, to server k, then

(55)
$$\lambda_{e,k} = \lambda \sum_{n \in BB} \pi_{\underline{n}}^* \left[\sum_{i=1}^{N-n} g_i \operatorname{AR}(\underline{n}, i, k) \right]$$

where $AR(\underline{n}, i, k)$ denotes the average number of customers accepted by server k if the batch contains i customers and the system is at state \underline{n} when the batch arrives. In (55), i ranges form 1 to N - n since it is necessary for a abatch not to contain more than N - n+ 1 if that batch is to be accepted. AR(n, i,

k) can be accepted as follows. First we define an indicator variable IV(k) as follows to indicate whether k is in I(n) of (49.b)

(56)
$$IV(k) = \begin{cases} 1, & \text{if } k \in I(n) \\ 0, & \text{elsewhere} \end{cases}$$

Using the notations defined in (49) we obtain

This can be explained as follows. When $i \le v(n) + IV(k)$ then the batch will definitely be accepted. Therefore in this case the average number of

customers accepted is simply $\sum_{j=1}^{i} j(\frac{i}{j}) r^{j} (1-r)^{i-j}$. However when

 $v(\underline{n})$ + $IV(k) \le i \le v(n)$ + $|I(\underline{n})|$ then the batch of size i could be rejected due to improper routing. The term $\sum q_X(i + j, R(\underline{n}), B(\underline{n}), I(\underline{n}))$ accounts for this effect and must be subtracted.

Once $L_{q,k}$ and $\lambda_{e,k}$ are obtained, the average waiting time of an accepted customer routed to server k, $W_{c,k}$, can be obtained from Little's formula as follows.

(58)
$$W_{C,k} = \frac{L_{q,k}}{\lambda_{e,k}}$$

This holds at steady state.

Then the average delay of an accepted customer routed to server k is

(59)
$$D_{C.k} = W_{C.k} + T$$

Let $D_{\mathbf{C}}$ denote the average delay of a customer, then

(60)
$$D_{c} = \sum_{k=1}^{m} D_{c,k} r_{k}$$

Although the derivation of $D_{\rm C}$ is done through (58) via Little's formula, the average batch delay $D_{\rm D}$ has to be obtained in a different manner. This is of course due to the variation in batch size. But more importantly, customers in the same batch can be routed to different servers. This complicates the derivation of $D_{\rm D}$.

Let $D_{b,\ell}$ denote the average daly of a batch of size ℓ . Then

(61)
$$D_{b,1} = \sum_{(\underline{a}, \underline{b}) \in AB_{1}^{\underline{b}}} p_{\underline{a},\underline{b}}^{*} \left\{ \sum_{k_{1} \in KK_{1}} [a_{k_{1}} + (b_{k_{1}} - 1)^{+} + 0.5]r_{k_{1}} \right\}$$

where

(62)
$$AB_{1} = \{(a, b) \mid 0 - a + b - N - 1, \sum_{j=1}^{m} (a_{j} + b_{j} - 1)^{+} - N - m\}$$

In (61), the 1 corresponds to 1 slot of service time while the 0.5 represents the average residual period. The residual period of a customer is measured from the instant of its arrival till the beginning of the next slot. Since arrival could occur anywhere within a slot, the average residual period turns out to be 0.5 slots. The KK_1 in (61) denotes the set of serves which guarantees the acceptance of the batch of size 1. Obviously, if there is at least one empty buffer in the shared region then $KK_1 = \{1, 2, ..., m\}$. If the shared buffer is already full, then KK_1 is simple the set of servers for which their reserved buffers are empty. For $D_{b,2}$ we have

(63)
$$D_{b,2} = \sum_{(a,b) \in AB} \pi_b p_{\underline{a},\underline{b}}^* \left\{ \sum_{(K_1, k_2) \notin KK_2} \max_{(a,b) \in KK_2} [a_{k_1} + (b_{k_1} + 1)^+ + d_{k_1} + 0.5, a_{k_2}] \right\}$$

$$a_{k_2} + (b_{k_2} - 1)^+ + d_{k_2} + 0.5]r_{k_1} r_{k_2}$$

where \mathbf{d}_{i} denotes the number of customers in the batch of size 2 which select server i and

(64)
$$AB_2 = \{(a, b) \mid 0 \le a + b \le N - 2, \sum_{j=1}^{m} (a_j + b_j - 1)^{+} \le N - m\}$$

(65)
$$KK_2 = \{(k_1, k_2) | \text{ servers } k_1 \text{ and } k_2 \text{ are chosen by the batch and } k_1 \text{ and } k_2 \text{ are chosen by the batch and } k_2 \text{ are chosen by } k_2 \text{$$

$$\sum_{i=1}^{m} (a_i + b_i + d_i - 1)^+ \le N - m$$

In general, we have

(66)
$$D_{b,\ell} = \sum_{(\underline{a},\underline{b})} \pi_{\underline{b}} p_{a,b}^{*} \{ \sum_{(k_{1},\ldots,k_{\ell})\in KK_{\ell}} \max_{k_{1}} [a_{k_{1}} + (b_{k_{1}} - 1)^{+} + d_{k_{1}} + 0.5,$$

...,
$$a_{k_{\ell}} + (b_{k_{\ell}} - 1)^{+} + d_{k_{\ell}} + 0.5] r_{k_{1}} ... r_{k_{\ell}}$$

where

(67)
$$AB_{\ell} = \{(a, b), 0 \le a + b \le N = \ell, \sum_{j=1}^{m} (a_j + b_j = 1)^+ \le N = m\}$$

and.

(68) $KK_{\ell} = \{(k_1, \ldots, k_{\ell}) \text{ servers } k_1, \ldots, k_{\ell} \text{ are chosen by the batch and } k_{\ell} \}$

$$\sum_{i=1}^{m} (a_i + b_i + d_i - 1)^+ \le N - m$$

Finally, the average batch delay is given by

(69)
$$D_{b} = \frac{1}{1 - B_{b}} \sum_{\ell=1}^{N} g_{\ell} D_{b,\ell}$$

Concerning delay analysis, most papers in the literature concentrate on customer delay, mainly for simplicity. For customer delay, relation between delay and throughput or input rate can be established usually through Little's formula. However, no such relation is guaranteed for batch or group delay. The reasons for the problem studied in this report are variation in batch size and routing effect. Similar phenomenon is also noticed by Whitt [6] and Halfin [7].

The system throughput S is given by

(70)
$$S = \sum_{b \in BB} \sum_{k=1}^{m} u(b_k) \pi_{\underline{b}}$$

where BB is defined in (5).

We have carried out extensive numerical calculations based on the results obtained here. Fig. 2.5 are part of them. In these examples we assume m = 3, N = 8 and $r_1 = r_2 = r_3 = 1/3$. Two batches size distributions are considered in each of these examples. One is the well known geometric distribution with p = 0.5 or α = 2. The other is specified by $g_1 = g_8 = 0.267$, $g_2 = g_7 = 0.133$, $g_3 = g_6 = 0.067$, and $g_4 = g_5 = 0.033$. Since the latter has a shape looks like a suspension bridge this distribution will be referred as SB for convenience.

In each figure, the curve labeled by G corresponds to geometric distribuion while the one labeled by SB corresponds to the suspension bridge distribution mentioned above. In addition to numerical calculations we also carry out simulations to support our analysis. We observe in these figures that the agreement between analysis and simulation is extrememly good.

Fig. 2 shows B_b vs λ . First, we observe that in both G and SB, B_b is an increasing function of λ . This is intuitively reasonable. As a matter of fact, B_b approaches 1.0 as λ approaches infinity. The rate of convergence depends on the actual distribution of the size. Second, we observe that SB distribution exhibits higher blocking probability than G. This is explainable. Since the size of a batch in SB is restricted to occur in the range between one and eight $g_1 = g_8 = 0.267$ is considerably higher than the rest. On the contrary, for geometric distribution with p = 0.5, $g_1 = 0.5$ and the

distribution is strictly decreasing in batch size. This then implies SB had higher blocking probability. Third the gap between G and SB closes up as λ increases. This is reasonable since as λ becomes big enough the system occupancy is high and most of the arrivals will be rejected regardless of their actual size distribution.

アンスの19 グライン・フィー・アントン・ジン・ 見たのなかから

Fig. 3 and 4 show D_C and D_b , respectively, versus λ . For G distribution we observe both D_C and D_b increase as λ increases. This is reasonable. However, for SB, D_C and D_b first decreases then increases as λ increases from zero. This can be explained as follows. As λ is low, almost every arrival can be accepted. Since the batch of size 8 has a good chance of entering the system when λ is small, the overall delay is high. When λ gradually increases, only short batches can enter the system thus the average delay of the accepted batches drops. As λ is high enough, system is full almost all the time thus the delay again is high. Notice that the disagreement between analysis and simulation is less than 5%.

Fig. 5 shows system throughput versus λ . Here we observe G has higher throughput. This phenomenon is consistent with Fig.2.

3. Extensions to Unequal Rate Sservers

Although routing has been considered in Section 2, the servers there are still assumed to be operating at the same speed. In practice, servers attached to the queue may run at different speeds. In order to move one step forward in the modeling of a packet switch, the purpose of this section is to extend the results to include servers possible running at different rates.

In this section we assume when a customer is routed to server k it will take the server (or transmitter) Δ_{k} seconds to complete its service for this customer. Due to the synchronous nature of servers, the behavior of the queue can be described on the basis of t-second slots where T is the greatest common divisor of Δ_{1} , ..., Δ_{m} . In addition to the random routing considered above we also consider another type of routing called idle-server-first (ISF) routing which is believed to offer better performance. In ISF routing, idle servers or servers with no customer waiting will be considered first upon the arrival of a packet. For both types of routing we have established state-transition equations, obtained state transition probabilities, and derived results such as blocking probabilities, delays, and throughputs, etc. Based on these results we have also carried out extensive numerical calculations. The validity of analysis has also been verified by computer simulations. Figures. 6-9 show part of the numerical results.

4. Conclusions

We have in this work studied the effect of routing and buffer sharing with minumum allocation on the behavior of a finite queue which receives batch Poisson inputs and provides multiple servers running synchronously at the same speed. The main contribution of this work is to obtain analytical results for system state probability, blocking, delay and throughput. The validity of analysis is not only verified by simulation but also supported by intuitive reasonings. The work reported here is an extension of [1].

We have also extended the model to include unequal rate servers.

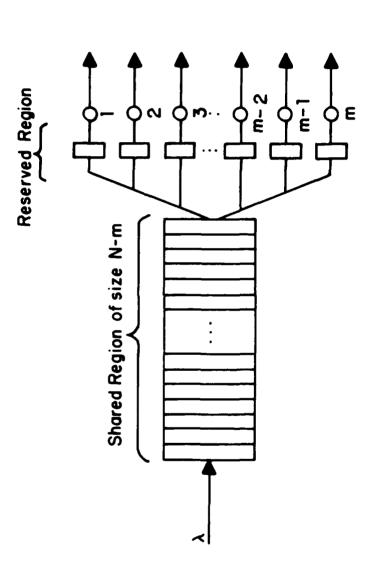
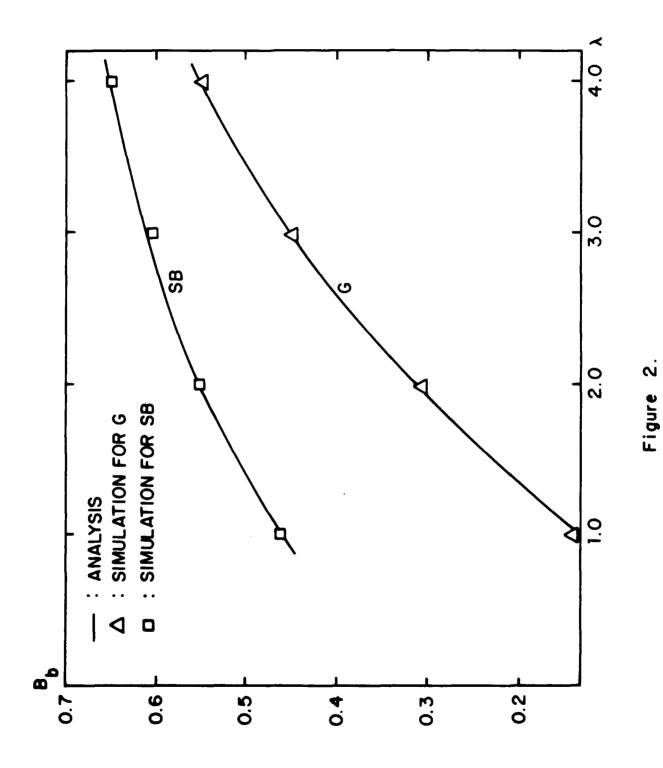


Figure 1.



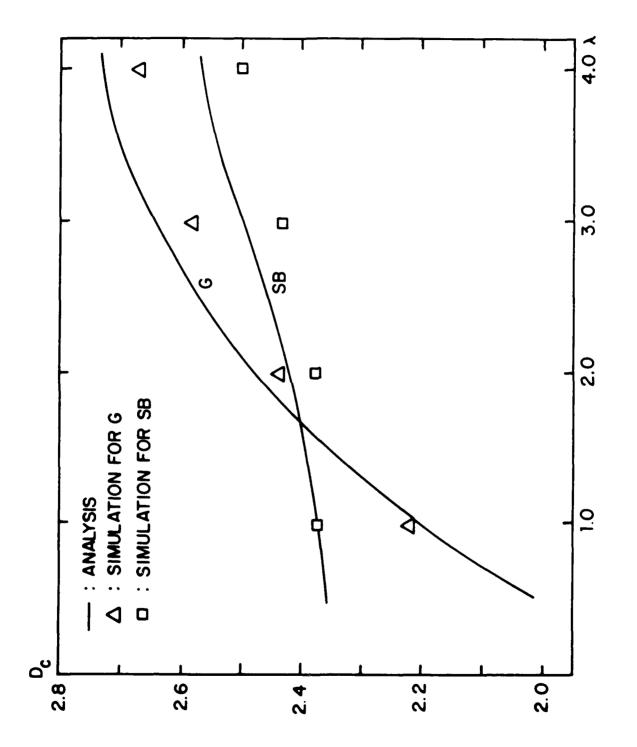
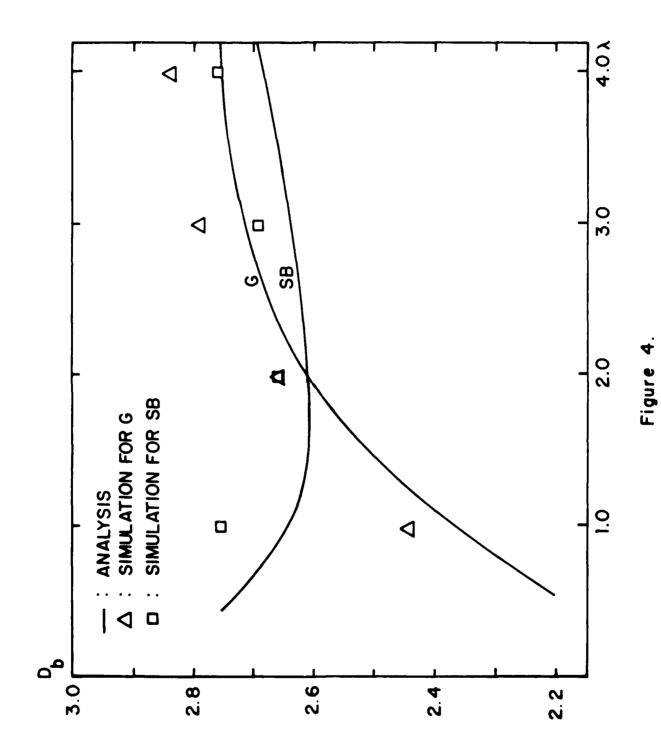
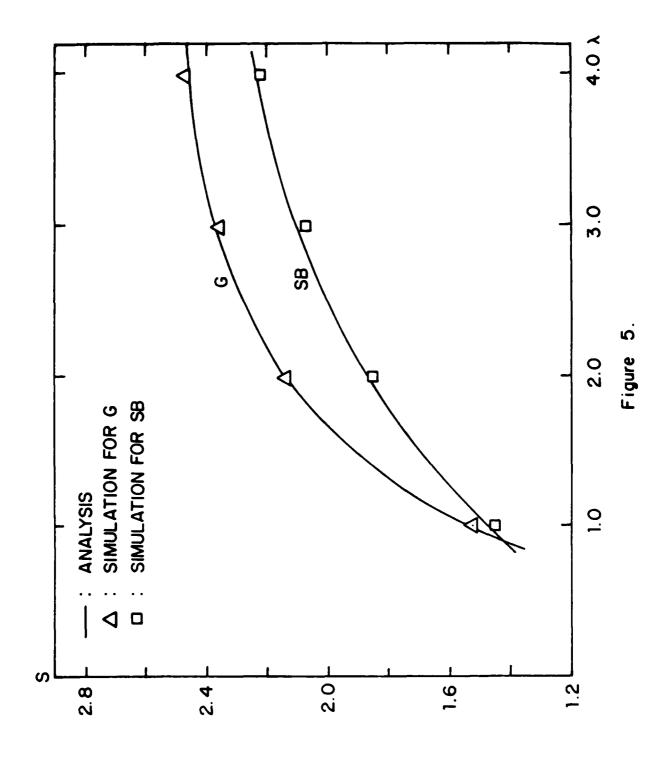


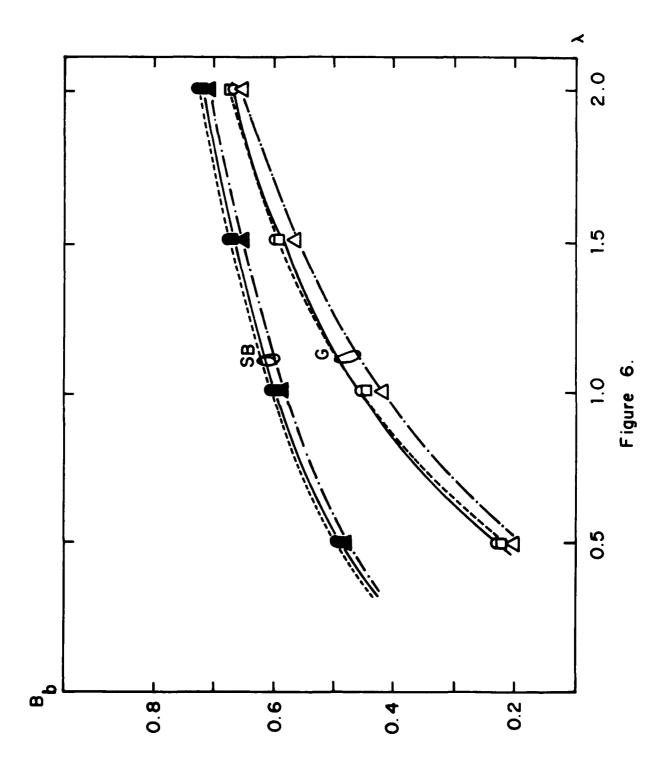
Figure 3.

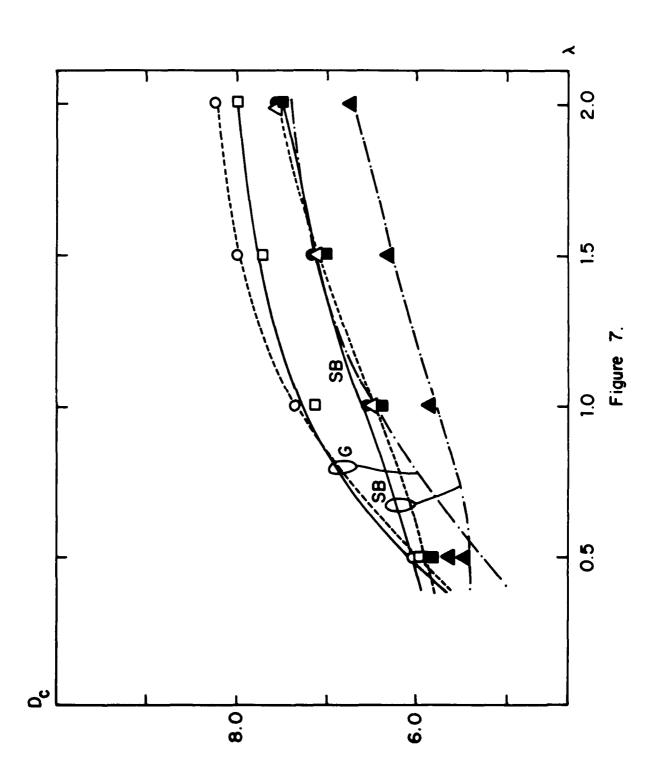


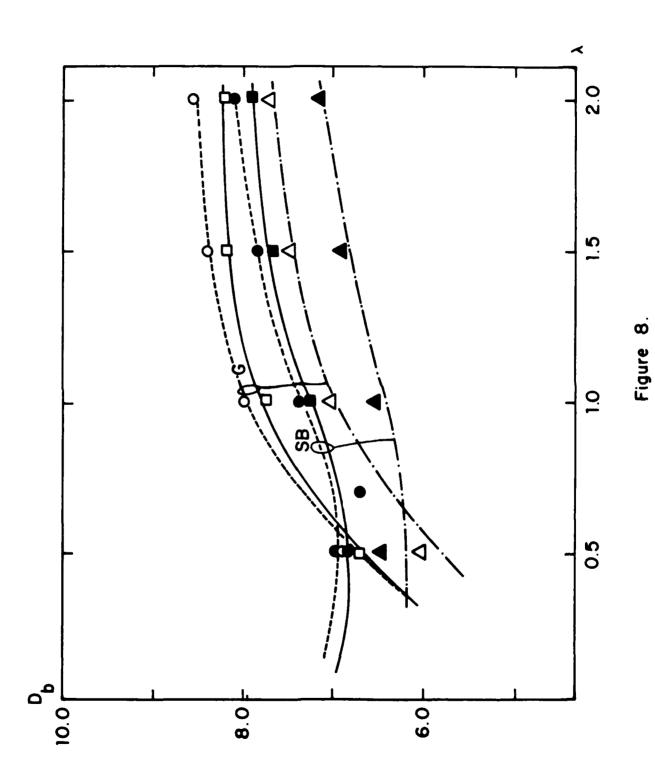


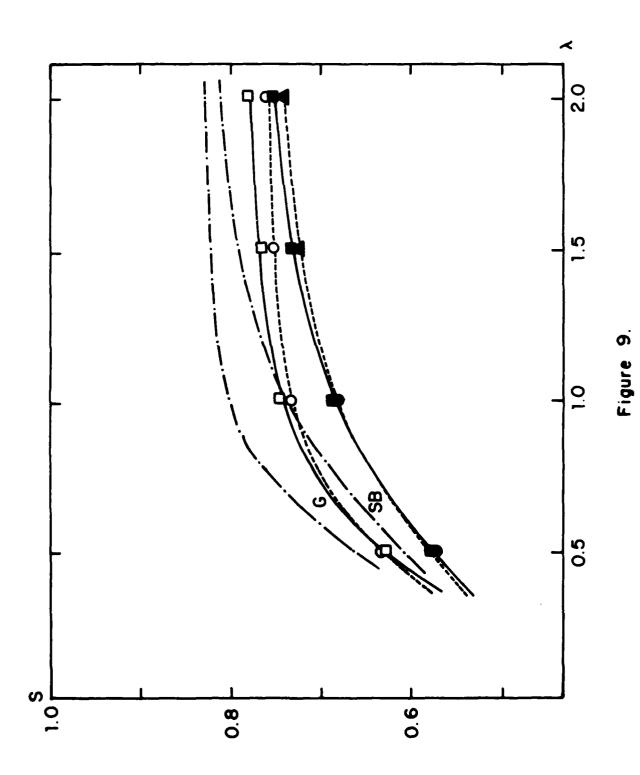
COURT CROSSOCIONS CONTROL OF

REPORT OF THE PROPERTY OF THE









DISTRIBUTION LIST

	NO. OF COPIES
Library, Code 0142 Naval Postgraduate School Monterey, CA 93943-5100	4
Director of Research Administration Code 012A Naval Postgraduate School Monterey, CA 93943-5100	1
Professor Jin-Fu Chang Code 62Cn Naval Postgraduate School Monterey, CA 93943-5100	5
Defense Technical Information Center Cameron Station Alexandria, VA 22314	2
Center for Naval Analyses 2000 N. Beauregard Street Alexandria, VA 22311	1